Authors:

1. Stien Heremans (corresponding author); Department of Earth and Environmental Sciences, KU Leuven; Celestijnenlaan 200E, 3001 Leuven, Belgium; stien.heremans@ees.kuleuven.be; +32 16 379398.
2. Jos Van Orshoven; Department of Earth and Environmental Sciences, KU Leuven; Celestijnenlaan 200E, 3001 Leuven, Belgium; jos.vanorshoven@sadl.kuleuven.be; +32 16 329740.

This work was supported by the Agency for Innovation by Science and Technology in Flanders (IWT) under Grant 101340.

Machine learning methods for sub-pixel land cover classification in the spatially heterogeneous region of Flanders (Belgium): a multi-criteria comparison

Stien Heremans and Jos Van Orshoven

Department of Earth and Environmental Sciences, KU Leuven (University of Leuven), Belgium.

Machine learning methods for sub-pixel land cover classification in the spatially heterogeneous region of Flanders (Belgium): a multi-criteria comparison

Until now, few research has addressed the use of machine learning methods for classification at the sub-pixel level. To close this knowledge gap, in this paper, six machine learning methods were compared for the specific task of sub-pixel land cover extraction in the spatially heterogeneous region of Flanders (Belgium). In addition to the classification accuracy at the pixel and the municipality level, three evaluation criteria reflecting the methods' ease-of-use were added to the comparison: the time needed for training, the number of meta-parameters and the minimum training set size. Robustness to changing training data was also included as the sixth evaluation criterion. Based on their scores for these six criteria, the machine learning methods were ranked according to three multi-criteria ranking scenarios. These ranking scenarios correspond to different decision making scenarios that differ in their weighting of the criteria. In general, no overall winner could be designated: no method performs best for all evaluation scenarios. However, when both the time available for preprocessing and the magnitude of the training dataset are unconstrained, Support Vector Machines clearly outperform the other methods.

Keywords: machine learning; land cover classification; sub-pixel

1 Introduction

Timely information on the distribution of land cover and land use types in time and space are crucial to a wide range of end-users, from the sub-national to the global scale (Kavzoglu and Colkesen 2009; Chapman et al. 2010; Colditz et al. 2011; Shao and Lunetta 2011; Rodriguez-Galiano et al. 2012). Many national and international organizations require up-to-date land cover information to analyze the structure and the functioning of different natural and artificial ecosystems and to monitor changes in these systems (Bontemps et al. 2011; Kaptué Tchuenté, Roujean, and De Jong 2011; Verburg, Neumann, and Nol 2011; Groisman and Bartalev 2007; Donlon et al. 2012). The continuous improvements of remote sensing techniques over the last decades have markedly reduced the costs of gathering this

information, turning them into a reliable alternative for more costly ground-based surveys. As a consequence, satellite and areal images have rapidly become important data sources in the field of land cover and land use assessment (Kavzoglu and Colkesen 2009; Rogan et al. 2008; Friedl, Brodley, and Strahler 1999).

Considering the currently available satellite missions, space-borne sensors with a moderate spatial resolution are most appropriate for regional/global land cover assessment, as they offer a global coverage of the Earth on a daily basis (Matsuoka et al. 2007; Shao and Lunetta 2011). Their short revisiting time however comes at the expense of spatial detail, resulting in an increased share of pixels that contain a mixture of different land cover types on the ground (mixed pixels) (Fisher 1997; Faivre and Fischer 1997). These mixed pixels pose a problem for hard land cover classification, due to the fact that their spectral characteristics do not originate from a single land cover class (Foody 1996). In the presence of mixed pixels, sub-pixel classification offers a valuable alternative to the traditional hard classification approach. It approaches the land surface as a set of area fraction images, one for each land cover class concerned (Verbeiren et al. 2008). The main objective in sub-pixel classification is to accurately estimate the fractional cover of each land cover class within each pixel. Sub-pixel classification can be considered as an intermediate between traditional hard *classification* and typical *regression*. On the one hand, it can be perceived as a constrained regression where the output fractions must stay within the [0, 1] range and must sum to one. At the same time however, the exhaustive area-covering set of classes indicate that it is an extension of the hard classification scheme. The modeling approaches used in sub-pixel classification should thus be able to handle the regression-like nature of the problem. Traditional classification methods like nearest neighbor and minimum distance classifiers do not qualify for this specific task.

Large scale land cover extraction from low resolution sensors comes with several specific requirements (DeFries and Chan 2000; Roosta and Saradjian 2007; Rogan et al. 2008; Carrão et al. 2010; Clark et al. 2010; Rodriguez-Galiano et al. 2012):

(i) At a large spatial scale, land cover classes are often spectrally heterogeneous with large interclass, rendering the creation of reliable endmembers difficult. Therefore, at this scale, algorithms that do not require endmembers are preferred;

(ii) To allow land cover classification over large areas, classification algorithms must be selected so as to minimize time-consuming human intervention and maximize automated procedures;

(iii) The selected classification method(s) should be able to handle the typical characteristics of real large-scale applications: noisy data, a complex data source and a small number of training observations relative to the study area.

Machine learning methods fulfill these requirements, thus rendering them particularly suited for the large scale classification exercise described in this paper.

State of the art machine learning algorithms like Support Vector Machines, Random Forests and Boosted Regression Trees are getting widely accepted in many remote sensing related studies and have shown robust and reliable regression and classification results. From recent literature, we can conclude that spectral unmixing and machine learning are becoming equally popular for performing land cover classification at the sub-pixel level (Bocco et al. 2012; Cortés, Girotto, and Margulis 2014; Fan and Deng 2014; Farook, Sivaraman, and Kesavaraj 2013; Reschke and Hüttich 2014; Schwieder et al. 2014; Wang, Shao, and Kennedy 2014; Zhang, Zhang, and Lin 2014; Benhadj et al. 2012).

Machine learning has proved to be at least as accurate as linear unmixing in multiple studies over the last decade (Berberoglu, Satir, and Atkinson 2009). In heterogeneous landscapes, we expect a strong prevalence of complex mixing patterns, which complicates the

specification of an appropriate (non-linear) spectral unmixing formula. Moreover, in areas characterized by a large intra-class variability, especially for the agricultural land cover classes, selecting appropriate endmembers is difficult. The combination of these two factors has lead us towards selecting machine learning methods instead of spectral unmixing for the sub-pixel land cover classification performed in this study. Among the most commonly used inductive classification algorithms nowadays are artificial neural networks (ANNs), classification and regression trees (CARTs) and support vector machines (SVMs) (Weng 2012).

Since the performance of machine learning algorithms has proved to be problem-dependent (Lu and Weng 2007; Caruana and Niculescu-Mizil 2006; Szuster, Chen, and Borger 2011), it is generally recommended to compare different candidate-algorithms in the context of a specific application (DeFries and Chan 2000; Sutton 2005; Kohavi and John 1997; Strobl 2009). Specifically for land cover and land use classification, Dixon and Candade (2008) recommended an extensive suitability analysis in order to utilize the remotely sensed data to its fullest extent. To the best of our knowledge, few scientific studies have compared the performance of machine learning algorithms in the specific context of fractional land cover area estimation (Schwieder et al. 2014; Liu and Wu 2005; Walton 2008). With this study, we want to encourage the remote sensing community to reflect on this issue and give it the attention it truly deserves. Therefore, our objectives are

(i) to systematically assess and to analyze the accuracy of six machine learning methods for predicting sub-pixel land cover fractions from multi-temporal MODIS NDVI composites in the spatially heterogeneous region of Flanders, Belgium;

(ii) to compare and rank these methods with regard to a range of evaluation criteria that reflect their suitability for sub-pixel classification in spatially heterogeneous regions.

There are multiple criteria for assessing the suitability of a classification algorithm in addition to its accuracy. Is it efficient in terms of processing time? Can it easily be automated? Are the results it produces stable or is it unacceptably sensitive to variations in the training data? The evaluation criteria used in this paper have been so selected as to reflect the tradeoffs that influence the design of an operational land cover classification procedure, e.g., the possibility for automation, the robustness with regard to the training set and the required computational resources. The possibility for automation is related to the number (and the type) of meta-parameters, as the time needed for meta-parameterization can take up a large proportion of the total processing.

2 Data

2.1 Study area and classification scheme

Flanders is the densely populated region in the north of Belgium with a total area of 13 521 km². Nearly a quarter of the area is urbanized and about half of the Flemish surface is occupied by agriculture (Van Daele et al. 2010). Open space and urban land are strongly interwoven in comparison to other western European countries. The fragmented character of the Flemish landscape was enhanced by the lack of a rigid spatial planning strategy (Poelmans and Van Rompaey 2009). As a result, Flanders is one of the most spatially heterogeneous regions in western Europe. Figure 1 displays the geographical location of Flanders within western Europe.

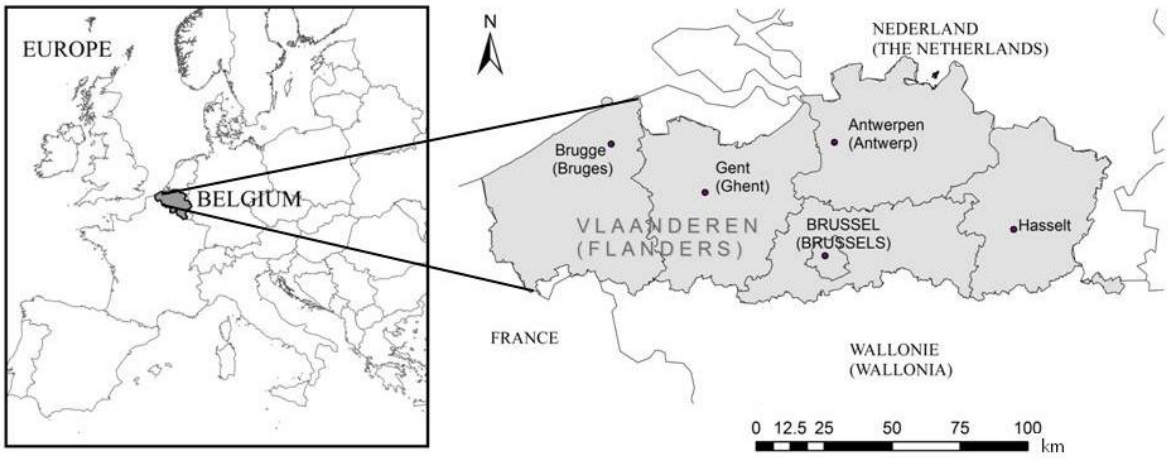

Figure 1: The location of Flanders in Europe.

Flanders was selected as the study area because of the fragmentation of its landscapes and the availability of a, yearly updated, high resolution crop coverage dataset that can be used as the main reference for training and validation of the machine learning methods. Tran, Julian, and de Beurs (2014) have demonstrated that sub-pixel classification was only advantageous compared to its per-pixel equivalent for heterogeneous landscapes.

2.2 Input data

The inputs to the machine learning methods in this paper consist of a time series of 8-day composites of MODIS (Moderate Resolution Imaging Spectroradiometer) NDVI (Normalized Difference Vegetation Index) covering the entire year 2010. NDVI is a normalized difference ratio of the red and infrared bands of the electromagnetic spectrum; high NDVI values are associated with healthy green vegetation (Jensen 2005). Time series of medium to coarse resolution NDVI values are one of the most used input datasets for land cover and crop classification at the regional to global scale (Lunetta et al. 2010; Wardlow, Egbert, and Kastens 2007; Roosta and Saradjian 2007; Shao and Lunetta 2011).

TERRA-MODIS data are freely available on the National Aeronautics and Space Administration (NASA) website. To create our input dataset, 46 8-daily NDVI images (MOD09Q1 product) covering the entire year 2010 were obtained through the USGS Global

Visualization Viewer (GloVis) service. These images are all atmospherically corrected NDVI composites at 500 m spatial resolution created by the constrained view-maximum value compositing (CV-MVC) algorithm of Huete et al. (2002). For any multitemporal analysis, the positional accuracy of the inputs is of paramount importance. This is particularly true for an approach such as the one presented here, which requires near perfect coregistration of the images in the time series stack. MODIS project scientists have reported 50 m geolocation accuracy at nadir (Wolfe et al., 2002), which is about one tenth of the pixel width and believed to be sufficiently low to assure reliable coregistration.

2.3 Reference data

For this study, a vector reference dataset covering the entire Flemish territory was created by combining the Flemish Single Parcel Registration dataset of 2010 with the European CORINE (Coordinated Information on the European Environment) Land Cover (CLC) dataset of 2006. The Single Parcel Registration (in Dutch: Eenmalige perceelsregistratie – EPR) is an annually updated geo-dataset created for governmental purposes that contains detailed information on most agricultural parcels present in the region (field boundaries, crop type, owner,…). It is the most detailed and up-to-date agricultural land cover dataset available for Flanders.

As the EPR contains information on agricultural parcels only (45% coverage of Flanders), the reference classification of the non-agricultural zones was extracted from the less detailed CORINE Land Cover 2006 dataset. Given that our study aims at classifying the land cover of 2010, there is a four year time lag between this CLC reference data and the inputs. In general however, non-agricultural areas under forest, grassland, infrastructure and water are more stable than agricultural parcels, which typically undergo annual crop rotations. The relative stability of the non-agricultural land is confirmed by Büttner et al. (2011) and

Hazeu et al. (2008), who estimated an overall change in the CORINE land cover classification between 2000 and 2006 of 1.25% for Europe and 1.62% for the Netherlands, respectively. When applying the same change rate to the period 2006-2010, we can safely presume that the four year time lag will not compromise the reliability of our reference data. The dynamic (agricultural) classes are covered by the annually updated EPR dataset and as such do not suffer from any time lag.

The CLC and the EPR datasets were combined by a simple overlay, with the CLC dataset as the base layer. For all agricultural parcels for which EPR data were available, the CLC base layer was replaced by this more detailed and up-to-date EPR data. After simplification of the legend, the resulting vectorial land cover dataset was used to extract a fraction raster for each cover class. These fraction rasters are the actual reference data used in this paper and their cells are spatially aligned with the MODIS pixels.

The final classification key contains 16 different land cover types, ten agricultural and six non-agricultural classes (Table 1).

Table 1: The land cover classification scheme used in this study and the areal share of each class in the Flemish territory in 2010.

Land cover class	**Proportion of area in Flanders(%)**
Grassland	18.25
Vegetables	1.66
Maize	9.54
Summer cereals	1.19
Winter cereals	4.23
Rapeseed	0.15
Potatoes	4.22
Beets	5.89
Orchard	1.08
Other agricultural LC	19.83
Built-up	23.32
Trees	8.00
Shrubs	0.88
Bare soil	0.06
Wetland	0.18
Water	1.14

3 Machine learning methods

Inductive machine learning refers to a class of computer algorithms that analyze data, extract patterns and then generalize these patterns to unseen data points through repeated learning from training instances. A large body of research is available that demonstrates the abilities of machine learning techniques to deal effectively with high-dimensional classification and regression problems. We recommend Gahegan (2003) for a detailed overview of the challenges and opportunities of machine learning for geo-information extraction purposes. This paper's line of reasoning convinced us to test a range of inductive machine learning methods for the challenging task of regional land cover assessment at the sub-pixel level. Six methods were selected, based on the recommendations of Caruana and Niculescu-Mizil (2006): the Multilayer Perceptron (MLP), Support Vector Regression (SVR) the Least-Squares Support Vector Machine (LS-SVM), Bagged Regression Trees (BaRT), Boosted Regression Trees (BoRT) and the Random Forest (RF). The following paragraphs contain a general description of these machine learning methods and their associated meta-parameters. Meta-parameters are parameters that govern the methods' behavior and prediction efficiency and they must be set by the user beforehand. An overview of all meta-parameter values tested can be found in Table 3. Table 2contains an overview of the software packages used for training the different machine learning methods.

Table 2: The software packages used for implementing the different machine learning methods in this study and a key reference for each package.

ML method	Software package used	Reference
Multilayer Perceptron	'Neural Network Toolbox' for Matlab	Beale, Hagan, and Demuth 2012
Support Vector Regression	'LibSVM'	Chang and Lin 2011
Least-Squares SVM	'StatLSSVM' for Matlab	De Brabanter et al. 2010
Bagged Regression Trees	'ipred' R package	Peters, Hothorn, and Hothorn 2012
Random Forest	'randomForest' R package	Liaw 2002
Boosted Regression Trees	'gbm' R package	Ridgeway 2007

3.1 Multilayer Perceptron (MLP)

A Multilayer Perceptron is a layered network that contains one input layer, one output layer and at least one hidden layer in between. These layers are composed of simple processing nodes and the nodes are interconnected between adjacent layers, but no interconnections between nodes in the same layer occur. Each interconnection carries an associated weight and each node passes the weighted sum of its inputs through an activation function that delivers an output value. This node output is either passed to all nodes in the subsequent layer or saved as the modeling output (for the output layer). The network is trained to correctly generate the output for unknown data points through an iterative learning process which adjusts the strength of the interconnection weights between the nodes (Rumelhart, Hinton, and Williams 1986).

The construction of a new MLP model is a challenging process, as a number of crucial meta-parameters have to be set up a priori (Hu and Weng 2009). The model's accuracy is mainly affected by three (sets of) meta-parameters: (1) the network architecture, (2) the learning algorithm and (3) the number of training iterations. Depending on the learning algorithm, 2-5 extra meta-parameters (the learning rate and the momentum term in the case of gradient descent learning) have to be set.

3.2 Support Vector Regression (SVR)

Support vector regression (SVR) refers to a family of regression models based on statistical learning theory (Vapnik 2005). The underlying principle is to fit a model based on a subset of the original training samples to predict a continuous response variable. Therefore, the inputs are mapped onto a high-dimensional space using an appropriate kernel function. In the new feature space, a linear model is then fitted that minimizes Vapnik's ε-insensitive loss and at the same time reduces the model complexity (Cherkassky and Ma 2004). Samples within a ε-

defined margin are considered to be well-represented and thus ignored. Samples outside this margin are penalized for their deviation (ε-insensitive loss) from the regression surface and as such determine its shape. A regularization parameter γ determines the trade-off between the model complexity and the ε-insensitive loss. The resulting linear regression function in a high dimensional (feature) space corresponds to nonlinear regression in the original input space. (Smola and Scholkopf 2004). The quality of an SVR model largely depends on the proper setting of its meta-parameters (Durbha, King, and Younan 2007; Kavzoglu and Colkesen 2009): (i) the kernel type and associated kernel parameter(s), (ii) the regularization parameter (γ) and (iii) the width of the error insensitive band (ε).

3.3 Least-Squares Support Vector Machine (LS-SVM) for regression

Least-squares support vector machines (LS-SVM) for regression (Suykens et al. 2002) are a special case of support vector regression where the inequality constraints have been replaced by equality constraints and the ε-insensitive loss function by squared loss. For a complete and detailed overview of LS-SVMs for both classification and regression purposes, we refer to Suykens et al. (2002). The meta-parameters associated with the LS-SVM for regression are (i) the kernel type and associated kernel parameter and (ii) the regularization parameter (γ).

3.4 Bagged Regression Trees (BaRT)

Tree-based models partition the input space into (multidimensional) rectangles, using a set of rules to identify regions with a homogeneous response to the input variables. Then, they fit the mean response for all (training) observations of that region to the entire region. The hierarchical structure of a tree means that the response to one input variable depends on values of inputs higher up in the tree, so interactions are automatically modeled. Individual regression trees have been identified as unstable learners highly sensitive to small perturbations in the training dataset (Breiman 1996): small changes in the training set can lead

to a very different output tree (Strobl 2009; Hastie, Tibshirani, and Friedman 2009). This instability introduces uncertainty in interpretation and limits predictive performance (Elith, Leathwick, and Hastie 2008). When multiple trees are grown from different training sets and their outputs averaged, a marked reduction of this variability can be realized (Breiman 1996). Models that implement the averaging technique are referred to as 'ensemble methods'. The concept of these ensemble methods has been discussed in the pattern recognition and machine learning communities for over two decads. The most prevailing ensemble approaches include bagging, boosting and their variations (Miao et al. 2012).

Bagging (bootstrap aggregating) is a relatively simple ensemble procedure that uses many bootstrap sets drawn with replacement from the original training dataset (Ismail and Mutanga 2010) and grows a regression tree from each bootstrap sample (Efron and Tibshirani 1993). The results of each individual tree are subsequently averaged to obtain the overall prediction. For creating a bagged regression tree ensemble, only one meta-parameter has to be set: the ensemble size (*k*).

3.5 Random Forest (RF)

Random Forests are a modified version of Bagged Regression Trees where the set of predictor variables is randomly restricted at each split (Prasad, Iverson, and Liaw 2006). This reduces the correlation between the individual trees, with the aim of improving the overall predictive power and the efficiency of the ensemble (Miao et al. 2012). Excluding some variables at each node also allows other input variables, that were otherwise outplayed by a stronger competitor, to enter the ensemble (Strobl 2009). Typically, a large number of trees are grown, hence the name 'random forest'. As more trees are added to the ensemble, RF models do not overfit but converge as they exhibit a bounded generalization error (Breiman 2001). RFs have been shown to be very efficient and to handle large datasets easily (Walton 2008). The RF

classifier needs the definition of three meta-parameters for generating a prediction model: (i) the ensemble size, (ii) the maximum size of the individual trees and (iii) the number of variables randomly selected at each node.

3.6 Boosted Regression Trees (BoRT)

In the bagging and random forest procedures described above, all individual trees are grown independently from one another. In contrast, boosting uses a forward stepwise procedure to iteratively fit trees to the training dataset and gradually increases emphasis on the poorly modeled samples (Hastie, Tibshirani, and Friedman 2009). Observations with a high error value in previous iterations receive a higher weight, thereby forcing the next tree to focus primarily on them. For a regression problem, boosting is equivalent to a functional gradient descent approach (Friedman 2002; Elith, Leathwick, and Hastie 2008). The final model can be interpreted as a linear combination of trees usually shrunk by a user-determined shrinkage factor to increase its performance (Elith, Leathwick, and Hastie 2008). The overall output value is again calculated by averaging the outputs of all the trees in the ensemble. Boosting was originally developed for weak learners, i.e. classifiers that are only slightly better than random guessing (Sutton 2005). Therefore, the tree size in a boosted ensemble is usually low, as opposed to bagging and random forests which mostly use (nearly) fully grown trees. Fitting of a boosted regression model requires the specification of three meta-parameters: (i) the ensemble size (k), (ii) the tree size (m) and (iii) a shrinkage factor (s).

The shrinkage factor is a weighting factor that controls the rate at which model complexity is increased. Smaller shrinkage values generally lead to more accurate models, but they require a larger number of trees to achieve the optimum (Hastie, Tibshirani, and Friedman 2009).

4 Methodological Approach

The six machine learning algorithms described above were applied in this study for estimating the sub-pixel fractions of 16 land cover classes (see Table 1). Some preprocessing was required, like temporal smoothing of the inputs and meta-parameterization of the algorithms. The main modeling part consisted of training the models and applying them to an unseen test set. Each algorithm was then evaluated with regard to eight performance criteria reflecting its predictive performance, ease of implementation (both in terms of preprocessing, training set requirements and training time) and robustness (with respect to the training data). Finally, the algorithms were ranked according to three multi-criteria evaluation scenarios.

A total of eight research steps were required to obtain the final rankings (see Figure 2):

(4.1) Temporal smoothing of the input data;

(4.2) Meta-parameterization of the algorithms;

(4.3) Model training and application;

(4.4) Aggregation of the outputs to the municipality level;

(4.5) Accuracy assessment;

(4.6) Assessing the impact of the training set size;

(4.7) Quantifying the robustness of the models to changing training sets;

(4.8) Performing a multi-criteria model ranking.

Below, you can find a detailed description of each step.

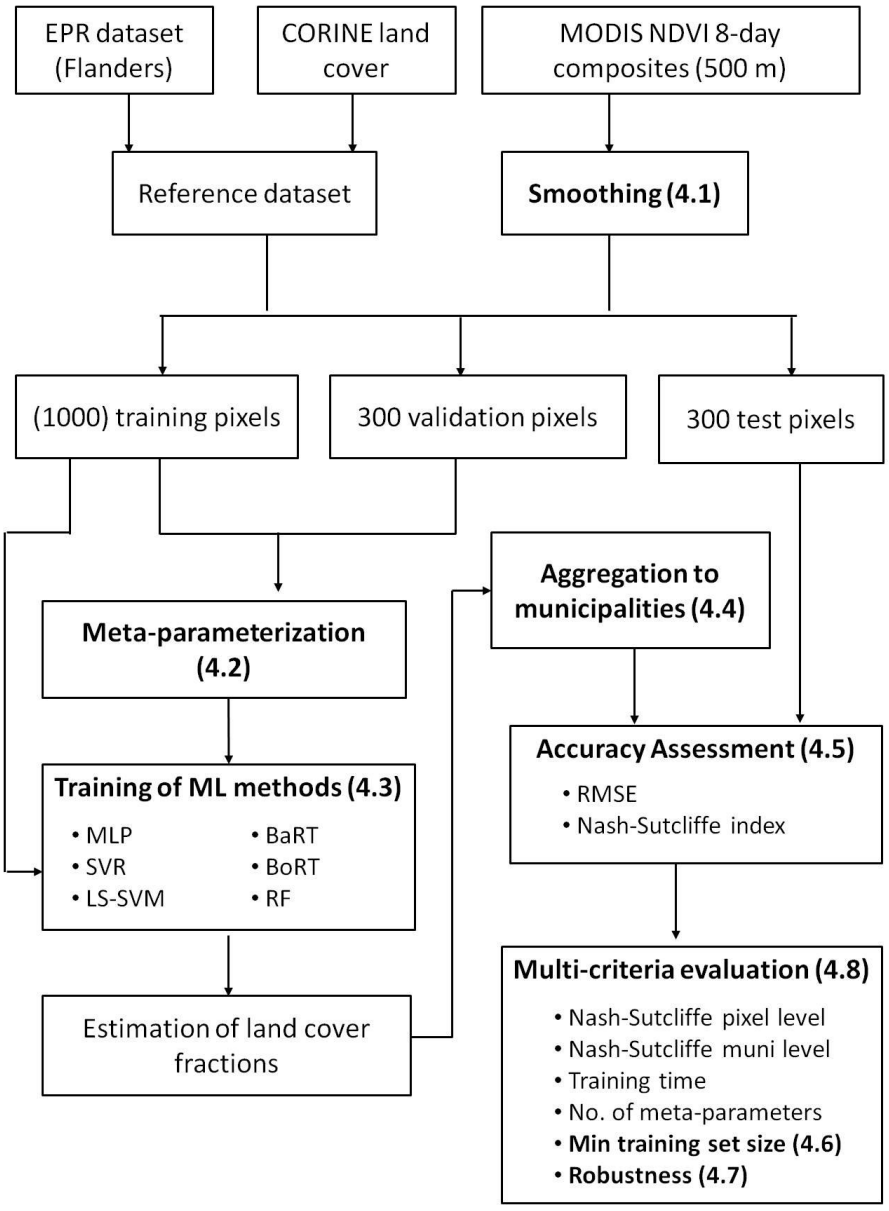

Figure 2: The methodological approach used to obtain a multi-criteria comparison of the six machine learning methods in this study.

4.1 Temporal smoothing of the inputs

After extraction of the study area (Flanders), the original set of 46 8-day MODIS images was stacked into a multi-temporal NDVI time series. A temporal smoothing algorithm was then applied to reduce the negative consequences of cloud/snow contamination. Based on the recommendations of Hird and McDermid (2009), the adapted Savitzky-Golay (SG) filter as implemented in the TIMESAT software package was selected for this smoothing operation. The SG filter was developed by Chen *et al.* (2004) and has proved to be both effective and

flexible on obtaining high-quality NDVI time series from noisy data. The convolution can be understood as a weighted moving average filter with the weights given as a polynomial of a certain degree. The general equation of the least-squares convolution for NDVI time series smoothing is

$$Y_j^* = \frac{\sum_{i=-m}^{i=m} C_i Y_{j+i}}{N} \quad (1)$$

where Y is the original NDVI value, Y^* is the smoothed NDVI value, C_i is the weight coefficient for the ith NDVI value and N is the size of the smoothing window. To conduct SG filtering, two meta-parameters need to be set: the half-width of the smoothing window (m) and the degree of the polynomial used for fitting. Usually, a larger window size produces a smoother result at the expense of flattening sharp peaks. According to Chen et al. (2004), the half-width of the smoothing window should be set between four and seven for extracting a reliable long-term trend without neglecting the variations inherent to any NDVI time series. Moreover, the degree of the polynomial is typically set between two and four. Based on these recommended ranges, a visual optimization of the two parameters was done. For our dataset, the optimal combination appeared to be a window half-width of five and a degree of the polynomial of two. To ensure the availability of this window for all composites, five extra observations were added at the beginning and the end of the series during smoothing. For a detailed description of the TIMESAT software and its implementation of the Savitzky-Golay filter, we refer to Jonsson and Eklundh (2004).

4.2 Meta-parameterization

Although machine learning methods have been reported to produce more accurate predictions than statistical methods, meta-parameter tuning is a major issue that largely affects their performance (Kavzoglu and Colkesen 2009; Gahegan 2003; Liu and Wu 2005; Shao et al.

2009; Shao and Lunetta 2012). Meta-parameters are specific characteristics associated with a machine learning method that determine how a model will be built and trained. To date there is no generally accepted method for selecting the optimal meta-parameter values, other than trial and error. Each of the machine learning methods in this study came with one or more meta-parameters that needed to be tuned (see section 0). As recommended in the machine learning community, the meta-parameterization step used an independent validation set. From the area-covering input and reference datasets, a training set of 1000 pixels and a validation set of 300 pixels were randomly extracted. This relatively high share of training data is supported by Verrelst et al. (2012), who concluded that machine learning methods for regression perform better with an increasing training to validation set ratio. Ten different combinations of training/validation samples were used to set the final meta-parameter values, to ensure a close approximation of the global optimum.

Table 3 gives an overview of the meta-parameters associated with each machine learning algorithm, the strategy used for their optimization and the range of tested values. Three optimization strategies can be distinguished: simple search, grid search and early stopping. A simple search strategy optimizes a certain parameter independently from all others. A grid search procedure on the other hand, approximates the optimum for a combination of two or more meta-parameters. Early stopping is considered as a special case of simple search, where the optimum (on the validation set) is located during the training phase.

Table 3: Search strategies used for the meta-parameterization of the ML algorithms. Note: the braces (}) group the parameters optimized in one grid search procedure. The asterisk (*) indicates parameters that were optimized for each class separately.

ML method	Meta-parameter	Optimization strategy	Tested values/types
MLP	No. hidden nodes	Simple search	[4; 6 ;8; 10; 12; 14]
	Learning algorithm	Previous research (Heremans and Van Orshoven 2011)	[gradient descent; conj. gradient; scaled conj. gradient; quasi-newton; Levenberg-Marquardt]
	No. iterations	Early stopping (see (Caruana, Lawrence, and Giles 2001))	[0-1000]
SVR	Kernel type	Simple search	[linear; Gaussian]
	Kernel parameter*	Grid search }	[0.5; 1; 2; 4; 8; 16; 32; 64]
	Regularization parameter*	Grid search }	[0.003906; 0.007813; 0.015625; 0.03125; 0.0625; 0.125; 0.25; 0.5; 1]
	Width of the error band*	Grid search }	[0.003906; 0.007813; 0.015625; 0.03125; 0.0625; 0.125; 0.25; 0.5]
LS-SVM	Kernel type	Simple search	[linear; polynomial; Gaussian]
	Kernel parameter*	Grid search }	[0.5; 1; 2; 4; 8; 16; 32; 64]
	Regularization parameter*	Grid search }	[0.003906; 0.007813; 0.015625; 0.03125; 0.0625; 0.125; 0.25; 0.5; 1]
BaRT	Ensemble size	Simple search	[25; 50; 75; 100]
RF	Ensemble size	Grid search }	[125; 250; 500; 750]
	Maximum tree size	Grid search }	[25; 50; 75]
	No. variable per split	Simple search	[3; 9; 15; 21]
BoRT	Ensemble size*	Early stopping	[1-10000]
	Tree size	Grid search }	[1; 2; 3]
	Shrinkage factor	Grid search }	[0.005; 0.01; 0.05; 0.1]

For the MLP, the meta-parameterization step focused primarily on identifying the appropriate number of nodes in the hidden layer. The impact of the learning algorithm was not addressed here, as the Levenberg-Marquardt algorithm had already been identified as the most accurate and the most stable alternative in a lengthy pre-study using non-smoothed MODIS NDVI time series that also contained simple gradient descent, conjugate gradient learning and quasi-Newton learning (Heremans and Van Orshoven 2011). Levenberg-Marquardt offers a good compromise between the convergence speed of Newton's learning and the limited storage requirements of gradient descent (Wilamowski et al. 2001).

4.3 Model training and application

Once the optimal values of the meta-parameters for each ML method were determined, models were trained with the optimal meta-parameter combinations. Ten independent sets of 1000 training samples were used to parameterize the primitive functions (node weights, individual trees,…) and the resulting models were then applied to an independent 300 pixel test set, that was not used in the meta-parameterization nor in the training phase. This independent test set ensures a fair assessment of the generalization capacity of the trained models.

4.4 Aggregation to the municipality level

All machine learning models were evaluated both at the pixel level and at the municipality level. To generate municipality-level accuracies, the models were applied to all pixels within the Flemish region (our study area). Then, the estimated per-pixel fractions, as well as their associated reference fractions, were aggregated to the municipality level for all 308 municipalities in Flanders. This aggregation consisted of a weighted summation with the weights set equal to the relative areal share of the pixel in each municipality.

4.5 Accuracy assessment

The accuracy of the models trained in step 4.3 was assessed at two levels of spatial aggregation: per pixel and per municipality. At both aggregation levels, the root mean square error (RMSE) was calculated as a first accuracy measure. In many regression studies, the coefficient of determination ($R2$) is also included as an indicator of a model's prediction accuracy. However, this coefficient does not take into account the fact that perfect prediction assumes an intercept equal to zero and a slope equal to one. Therefore, in this paper, the more traditional $R2$ is substituted by the Nash-Sutcliffe (NS) index, which does take these two

constraints into account. The NS index is a normalized statistic that determines the relative magnitude of the residual variance (noise) compared to the measured data variance (information). NS indicates how well the plot of observed versus simulated data fits the 1:1 line (Nash and Sutcliffe 1970). It is calculated as

$$\mathrm{NS} = 1 - \left(\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}\right) \quad (2)$$

where $\hat{y}_i$ is the reference value and y_i the estimated value for the ith data point, $\bar{y}$ refers to the average of the reference values and n to the total number of data points in the test set. A perfect match between the estimated and the reference values will result in an NS value of 1. For unbiased models, the NS value is between 0 and 1. Biased models can also produce negative values of this indicator.

4.6 The impact of the training set size

Many studies have analyzed the effect of training set size on the accuracy of a variety of classification methods. In general, there was a positive relation between the number of training samples and the classification accuracy (Arora and Foody 1997; Foody, McCulloch, and Yates 1995; Pal and Mather 2003; Zhuang et al. 1994; Huang, Davis, and Townshend 2002). However, the acquisition of large training sets is both costly and time-consuming (Jackson and Landgrebe 2001). Therefore, we prefer algorithms that can build accurate prediction models from small sets of training data. The nature and complexity of the classification algorithm may have a marked impact on the minimum training set size required for accurate prediction (Kavzoglu and Mather 2003). Over the last decades, several studies have shown that machine learning techniques like neural networks, support vector machines and decision trees may often be able to achieve a high accuracy with a relatively small number of training samples (Durbha, King, and Younan 2007). Their non-parametric nature seems to reduce the requirement for a full and representative description of the function to be

estimated in feature space (Foody 2004).

For this paper, meta-parameterized ML models were trained at seven different training set sizes. Training sets contained 10, 50, 100, 250, 500, 1000 and 2500 pixels which respectively corresponds to an area equal to 0.016, 0.080, 0.16, 0.40, 0.80, 1.60 and 4.00 percent of the total study area. Since an area-covering reference dataset was available, the training sets were randomly sampled with exclusion of the fixed set of 300 test pixels. This allowed to estimate the minimum training set size needed for accurate area fraction estimation.

4.7 Robustness to changing training sets

Machine learning methods and other data-driven algorithms can be strongly affected by (subtle) differences in the training set (Rogan et al. 2008; Huang, Davis, and Townshend 2002). This instability has negative implications for the generalization capacity of the resulting prediction models. Therefore, it is important to select algorithms that are as robust as possible with regard to changes in the training set. The availability of an area-covering reference dataset allowed us to train each ML algorithm with 10 randomly selected training sets of 1000 pixels. The standard deviation associated with these 10 random repeats was used as an indicator of robustness with regard to the training data. The indicator values are inversely related to the algorithms' robustness.

4.8 Multi-criteria evaluation and ranking

Understanding the trade-offs associated with each classifier is important when trying to choose the optimal method for a given area and application. The 'best' or 'most appropriate' classifier is ultimately a subjective decision that depends on the specific objectives of the application, the characteristics of the data used and the available resources (Szuster, Chen, and Borger 2011; DeFries and Chan 2000; Lu and Weng 2007). It is not our intention to

formulate specific guidelines for the selection of a sub-pixel classification method, but merely to inform potential users about some criteria and tradeoffs that may influence this decision in an operational setting. Most certainly, these tradeoffs include: (i) the estimation accuracy, (ii) the computational resources needed and (iii) the ability to automate the process (DeFries and Chan 2000).

In this study, a total of six criteria were evaluated: (1) the accuracy (RMSE and Nash-Sutcliffe index) at the pixel and (2) at the municipality level, (3) the time required for training, (4) the number of meta-parameters to be set, (5) the minimal number of training pixels needed to guarantee an acceptable performance and (6) the robustness to changing training sets.

We have defined three ranking scenarios that reflect three different decision making scenarios. In the first, most general scenario, each performance criterion was simply given the same weight. The algorithms were first ranked according to each individual performance criterion and then these mono-criterion rankings were averaged to obtain the final multi-criteria ranking. Scenario 2 excludes the number of meta-parameters and the minimum training set size from the comparison, because they constitute of a limited number of discrete values and are as such unfit for linear scaling. In this scenario, the indicator values were first linearly scaled to the [0-1] range, where 0 indicates the worst performance and 1 the best. Then, these scaled values were averaged and a multi-criteria ranking derived from these averages. The scaling approach seems more fair than ranking averages (scenario 1), since it preserves relative differences in performance between the algorithms.

Scenario 3 is identical to scenario 2, with exclusion of the training time. Considering the computational resources available to most users nowadays, training time did not seem as crucial for model selection as the other criteria. In all three scenarios, only the NS index was included as an indicator for model accuracy, as it is strongly correlated to the RMSE and we wanted to avoid a double weighting of the accuracy.

5 Results

The results section focuses on the three least straightforward evaluation criteria: the meta-parameterization of the ML methods (5.1), their response to the training set size (5.2) and their robustness to changing training data (5.3). At the end of the section (5.4), the overall rankings according to the three ranking scenarios are displayed. All models trained in this paper have 46 inputs (MODIS NDVI composites) and 16 outputs (land cover fractions).

5.1 Meta-parameterization

The meta-parameterization step cannot be overlooked in inductive machine learning, since it is known to be one of the main determining factors for the final model accuracy. It is not our objective to elaborate on the effect of the individual parameters, but merely to illustrate the importance of this pre-processing step for a fair comparison of the algorithms. For the support vector methods (SVR and LS-SVM) and the tree ensemble methods (BaRT, BoRT and RF), a separate model was trained for each of the 16 land cover classes. This implies that some meta-parameters associated with these models were optimized for each class separately. For reasons of clarity, only the results for the meta-parameters optimized over all classes are included in Table 4 and visualized in Figure 3. In the model comparison and ranking section (5.4), all indicators refer to models trained with these optimal meta-parameter values.

Table 4: Optimal meta-parameter values for the machine learning methods used

ML method	Meta-parameter	Optimal value/type
MLP	No. hidden nodes	12
	Learning algorithm	Levenberg-Marquardt
	No. iterations	Variable (early stopping)
SVR	Kernel type	Gaussian
LS-SVM	Kernel type	Gaussian
BaRT	Ensemble size	75
RF	Ensemble size	250
	Maximum tree size	75
	No. variable per split	21
BoRT	Ensemble size	Optimum (pixel)/ Optimum x 8 (mun)

	Tree size	3
	Shrinkage factor	0.01

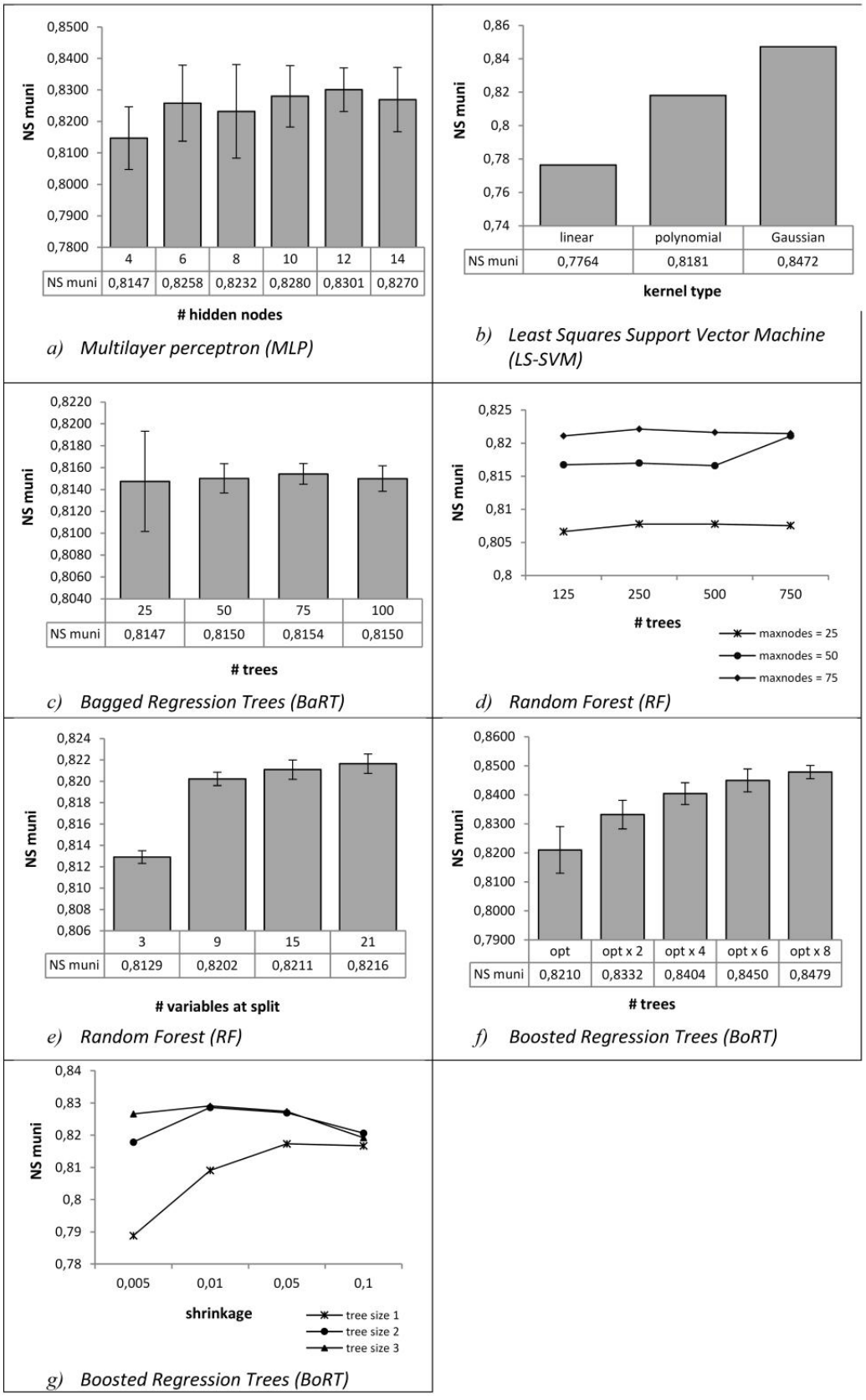

Figure 3: Visualization of some of the meta-parameterization results displayed in Table 4. 'NS muni' refers to the Nash-Sutcliffe index at the municipality level. The different subfigures display the meta-parameterization of the MLP (a), the LS-SVM (b), the BaRT (c), the RF (d) and the BoRT (f-g)

The main meta-parameter tested for the MLP was the number of nodes in the hidden layer. Figure 3(*a*) shows that, although the impact of this parameter is limited in the range of 10-14 nodes, an optimum can be identified at 12 hidden nodes. As explained in section 4.2, all MLPs were trained with the Levenberg-Marquardt algorithm, as implemented in Matlab's Neural Network Toolbox.

For Support Vector Regression, as well as Least Squares Support Vector Machines, the main meta-parameter to be optimized is the kernel type. Three kernel types were compared: a linear, a polynomial and a Gaussian kernel. The associated kernel parameter(s) and the regularization parameter were automatically optimized using the grid search procedures as specified in the LS-SVM package for Matlab or the LibSVM software for LS-SVM and SVR respectively. For both methods, the Gaussian kernel markedly outperforms the polynomial and the linear one (see Figure 3(*b*) for LS-SVM).

For Bagged Regression Trees, only the number of trees in the ensemble was optimized. Beyond 25 trees, only a marginal increase in accuracy can be noted (Figure 3*c*)). This confirms the findings of Prasad et al. (2006) and of DeFries and Chan (2000) who found that classification accuracy does not increase significantly beyond 25 and 50 trees respectively. However, in our experiment, the variance in accuracy decreases with the number of trees until a minimum is reached at an ensemble size of 75. Since robustness is also important for obtaining reliable results, the optimal number of trees was set at 75.

For the Random Forest method, the effect of three parameters was assessed: the ensemble size, the maximum tree size and number of randomly selected variables per node. The ensemble size and the maximum tree size were optimized in one single grid search procedure. Figure 3(*d*) shows that the ensemble size has almost no effect on the prediction accuracy. Our findings thereby confirm those obtained by Rodriguez-Galiano et al. (2012)

and Peters et al. (2007), who found that the generalization capacity of their RFs did not increase markedly after 100 trees. The maximum tree size on the other hand, does influence the model accuracy, with an optimum at 75 nodes. At the optimum of 125 trees and 75 nodes per tree, the number of random variables was optimized by a simple search. Prasad et al. (2006) used 1/3 of the total number of inputs, which corresponds to the default setting for the 'RandomForest' package. Rodriguez-Galiano et al. (2012) found that once the error starts to converge, the number of random variables per node hardly influences the generalization accuracy. For our dataset, convergence already starts at nine variables, which is markedly less than 1/3 of the number of inputs (n = 46).

Finally, the Boosted Regression Trees method required the setting of three meta-parameters: the ensemble size, the shrinkage factor and the interaction depth. The shrinkage factor and the interaction depth were optimized in a single grid search, while a software-embedded early stopping procedure ensured the optimization of the ensemble size at each grid point. Theoretically, very small values of the shrinkage factor guarantee the best generalization capacity, but at the expense of training time (Elith, Leathwick, and Hastie 2008). High shrinkage factors on the other hand are prone to the risk of overfitting. Therefore, intermediate values between 0.01 and 0.05 are generally recommended. An interaction depth (individual tree size) between 4 and 8 nodes has been identified as optimal for boosting (Hastie, Tibshirani, and Friedman 2009). This contrasts the findings for Bagged Regression Trees, which generally require large – mostly fully grown – trees. The grid search procedure for the BoRT model yielded an optimal shrinkage factor of 0.01 and an optimal interaction depth of three, which correspond to the default values of the R 'gbm' package. For the ensemble size, the optimal values obtained by early stopping only reflect the pixel level and as such does not always correspond to the optimum at the municipality level. Therefore, BoRTs were also trained with ensemble sizes exceeding the early stopping optimum. Figure

3(*g*) shows that the municipality-level accuracy continues to increase with the number of trees, far beyond the pixel-level optimum. To obtain optimal accuracy at the municipality level, training has to be prolonged until at least eight times the pixel-level optimum.

5.2 Impact of the training set size

From Figure 4, we can see that the relative accuracies of the six methods are strongly influenced by the size of the training set (number of training points). For 100 training points and less, Boosted Regression Trees performs best, followed by Random Forest. Even with the smallest training set (only 10 points), BoRT is capable of obtaining a municipality-level NS index (NS muni) of 0.79, while most other methods don't even reach 0.50. The Boosted Regression Trees method remains the most accurate method as the training set size increases but from 500 training pixels onward, Random Forest is outperformed by Support Vector Machines (both SVR and LS-SVM). A similar pattern can be identified with regard to the relative accuracies of the Multilayer Perceptron and Bagged Regression Trees. Below 500 training instances BaRT outperforms MLP, but for larger training sets the MLP becomes the better choice. At 2500 training observations, the saturation point does not seem to be reached for most algorithms. Still, as larger training samples are practically unachievable for most regions and strongly increase the computational burden, we have decided not to test sample sizes beyond 2500.

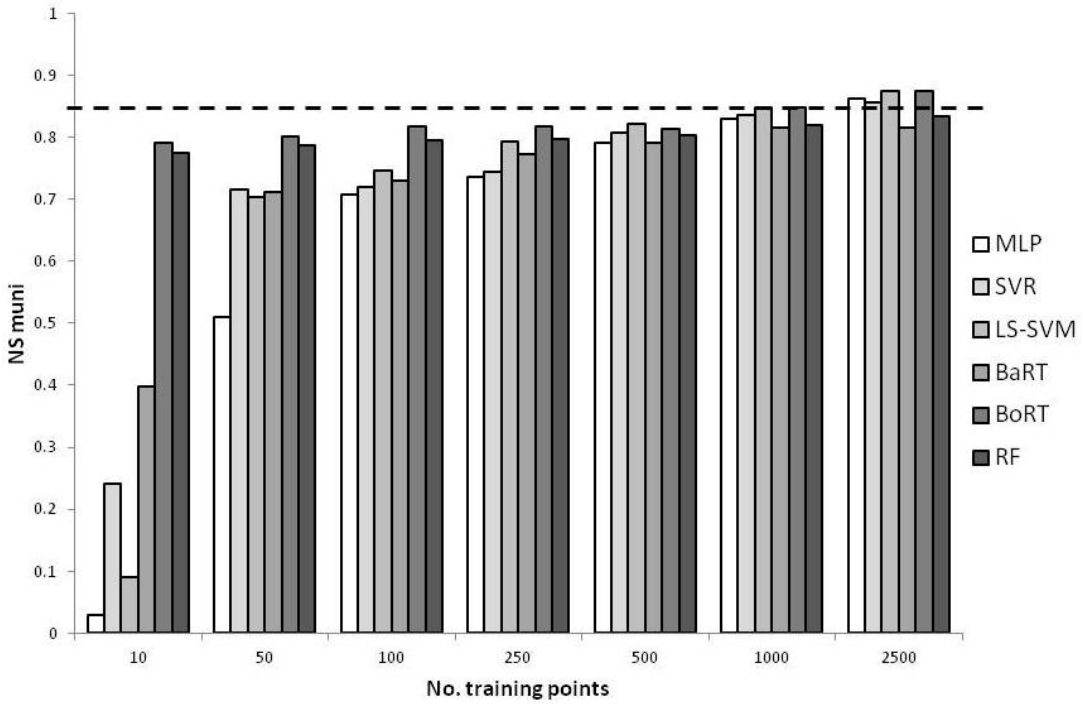

Figure 4: The municipality-level Nash-Sutcliffe index (NS muni) at increasing training set sizes for the six machine learning methods: MLP, SVR, LS-SVM, BaRT, RF and BoRT. The dashed line is the 0.80-line that distinguishes between non-sufficient and sufficient accuracy.

The response of a machine learning method to the size of the training set is an important criterion when selecting the most 'appropriate' method for a real-life application. That's why we aimed at developing a methodology that summarizes the observed patterns into one single measure. This measure could then be included in the multi-criteria ranking of the methods, as discussed in section 5.4. Based on the patterns identified from Figure 4, a Nash-Sutcliffe index at the municipality level of 0.80 seemed to be an appropriate threshold for distinguishing between a sufficient and a non-sufficient sub-pixel classification accuracy. At the lowest training set sizes, no method exceeds this threshold while at the highest training set size, all methods surpass it. Moreover, there is considerable variability in the point at which the different methods intersect the 'NS = 0.80-line'. Hence, the location of this intersection point will in this paper be used as a proxy for the complex response to the size of the training set and as such incorporated into the ranking scenarios of section 5.4.

5.3 Robustness to changing training sets

To ensure reliable predictions when reference data are hard to obtain, the accuracy of any modeling approach should be robust to changes in the reference set. As we have access to an area-covering reference dataset, we were able to explicitly test the response of the different machine learning methods to changing training sets. Figure 5 displays the standard deviation of the Nash-Sutcliffe index at the municipality level as obtained from 10 models, each trained with a different set of 1000 randomly selected training points.

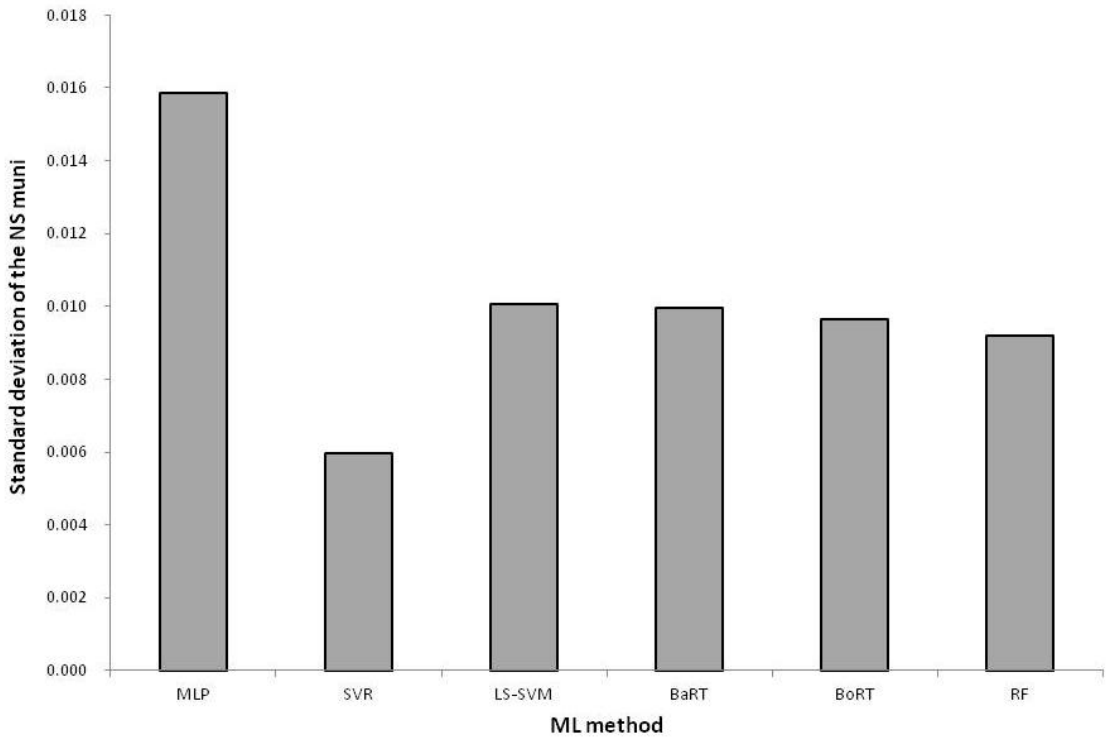

Figure 5: The standard deviation of the municipality-level Nash-Sutcliffe index (NS muni) obtained when training the methods with 10 different sets of 1000 training points. Robustness is inversely related with this standard deviation.

Figure 5 shows that all machine learning methods are relatively robust to changes in the training set, with standard deviations below 0.016 on average NS indices above 0.83. Thus,1000 training point are sufficient to accurately represent the variability in our study area. The most robust method is the SVR, while the MLP is the bottom of class for this indicator. The four other methods are equally robust. We may thus conclude that for our specific

application, robustness does not seem to be a majorly important consideration.

5.4 Multi-criteria evaluation and ranking

Table 5 summarizes the values for all the evaluation criteria included in this paper. The Nash-Sutcliffe (NS) index and the RMSE were used to represent the classification accuracy, both at the pixel and at the municipality level. The number of meta-parameters is a fixed characteristic of each ML method (see section 3). The intersection point with the NS = 0.80-line at the municipality level was used to summarize the effect of the training set size. Also, the training time and the robustness of the models to changing training sets were included.

Table 5: Overview of the performance criteria used to evaluate the machine learning methods. For each performance criterion, the highest performance is italic and underlined. In case of a tie, all (more or less) equally performing methods were indicated.

Indicator	MLP	SVR	LS-SVM	BaRT	RF	BoRT
NS index pixel level	0.456	0.476	*0.490*	0.453	0.471	0.456
RMSE pixel level	0.1128	0.1106	*0.1092*	0.1130	0.1110	0.1128
NS index muni level	0.8301	0.836	*0.847*	0.815	0.820	*0.848*
RMSE muni level	0.0422	0.0414	*0.0400*	0.0440	0.0434	*0.0399*
Training time (s)	22	*5*	*6*	202	27	400
No. meta-parameters	3-8	3-4	2-3	*1*	3	2
Min. training set size	1000	500	500	1000	250	*50*
Robustness (stdev)	0.0159	*0.0060*	0.0101	0.0097	0.0092	0.0097

The table above clearly shows that no ML method scores best for all the evaluation criteria. When taking only the accuracy for a single set of 1000 training pixels into account, Least-Squares Support Vector Machines seems to be the optimal choice, as it displays the highest accuracy both at the pixel and at the municipality level. However, in data-poor circumstances, Random Forests and Boosted Regression Trees outperform Support Vector Machines. When the time for pre-processing is limited, for example in near real-time estimations, the number of meta-parameters will become a crucial factor and Bagged Regression Trees may become the better choice. Training time does not seem an important criterion in our study, as all

methods can be trained in less than ten minutes. However, in studies carried out with limited computational resources, it could become an extra consideration. With regard to the robustness of the models, LS-SVMs do not perform as good as SVR.

Table 6 summarizes the ranks of the ML methods for the different performance scenarios. Scenario 1 includes all six performance criteria (Table 5), scenario 2 excludes the number of parameters and the minimum size of the training set and scenario 3 contains only the NS index at the pixel and the municipality level and the robustness to changing training sets. The accuracy (NS index) at the municipality level is included as a fourth 'scenario', as this is the criterion that actually needs to be optimized. The other criteria reflect the different boundary conditions this accuracy is subject to.

Table 6: Ranking of the machine learning methods according to the three ranking scenarios. The rank of best performing method(s) is italic and underlined.

Ranking scenario	**MLP**	**SVR**	**LS-SVM**	**BaRT**	**RF**	**BoRT**
Scenario 1	6	*1*	*1*	5	3	3
Scenario 2	6	2	*1*	5	4	3
Scenario 3	5	2	*1*	6	3	4
Accuracy municipality	4	3	*1*	6	5	*1*

All evaluation scenarios indicate the LS-SVM as the machine learning method with the best overall performance, closely followed by SVR. Support Vector Machines thus come out as the best choice for sub-pixel land cover classification in Flanders from a time series of MODIS NDVI. Bagged Regression Trees and Multilayer Perceptrons are situated at the lower end of the performance spectrum. Boosted Regression Trees are capable of obtaining good classification accuracies at the municipality level, but their average performance at the pixel level and their relatively long training time lower their overall ranking in most scenarios.

6 Discussion

6.1 Ranking of the methods

Various studies have demonstrated that no one machine learning method works best in all situations. This suggests that it may be wise to develop an understanding about which type of method works well for which type of dataset. This study can be considered as a small contribution to this quest, in trying to illuminate the behavior (accuracy and ease-of-use) of a set of six popular machine learning algorithms for the specific case of sub-pixel classification. Sub-pixel classification is both an important and challenging task in remote sensing (DeFries and Chan 2000; Rogan et al. 2008). The performance of the six algorithms included in this study was compared in a multi-criteria framework, including predictive power, robustness and ease-of-use. The choice for machine learning methods was pragmatic, as the heterogeneous nature of our study area – with high intraclass variability and a high prevalence of complex mixing patterns – did not allow a straightforward implementation of spectral unmixing. A direct comparison of our results to those of spectral unmixing was therefore not feasible. As the main objective of this study was to compare the performance of machine learning methods, this needs not be experienced as a major shortcoming. We have clearly demonstrated the strong performance of the selected machine learning methods in the framework of sub-pixel land cover classification at the regional scale (NS muni > 0.80). We feel that their merits (both in terms of performance and ease-of-use) were such that they can be recommended for similar studies, regardless of their relative position with respect to (linear) spectral unmixing.

The results reveal that all algorithms were able to predict the general land cover patterns, while Support Vector Machines and Boosted Regression Trees outperformed the other methods with regard to their prediction accuracy at the municipality level. Depending on the evaluation scenario used, a different ranking of the methods was obtained. As each

real-life application will come with its own unique set of boundary conditions, the weighting of the criteria are problem-dependent. Adaptation of the weights may lead to a different ranking and thus the preferred machine learning method will also depend on the specifications of the application. Despite its popularity in many application areas, the extensive meta-parameterization requirements are often mentioned as one of the major drawbacks of the Multilayer Perceptron (Liu and Wu 2005; Kavzoglu and Mather 2003; Verrelst et al. 2012; Berberoglu, Satir, and Atkinson 2009). Different potential combinations of model parameters lead to a large number of trials that have to be computed and summarized (Shao and Lunetta 2012). The results presented in this paper also indicate that for sub-pixel classification in Flanders, the MLP is outperformed on many fronts by other, more recent machine learning methods. This combination of suboptimal performance and extensive preprocessing needs leads us to not recommend the use of Multilayer Perceptrons for the specific application of sub-pixel land cover classification in heterogeneous landscapes in favor of Support Vector Regression (both LS-SVM and SVR) and Boosted Regression Trees.

In several land cover classification studies over the last decades, the empirical performance of Support Vector Machines (SVMs) was found to be competitive with the best available alternatives (Mountrakis, Im, and Ogole 2011; Camps-Valls et al. 2004; Huang, Davis, and Townshend 2002; Dixon and Candade 2008; Kavzoglu and Colkesen 2009; Shao and Lunetta 2012; Camps-Valls and Bruzzone 2005). The good performance of SVMs has been attributed to the low number of model parameters that have to be optimized, thereby reducing the possibility of overfitting (Brown, Gunn, and Lewis 1999). Moreover, the SVM is firmly grounded in statistical theory and since it uses quadratic programming, it can always locate the global minimum, whereas alternative algorithms tend to end up in local solutions that depend on the randomly selected starting point (Durbha, King, and Younan 2007; Mountrakis, Im, and Ogole 2011).

In many studies involving Classification and Regression Trees (CART), boosting has been identified as a very promising ensemble tree approach that often ranks among the best. Based on a comparative study including 57 publically available datasets, Banfield et al. (2007) showed that boosting and Random Forest performed significantly better than bagging. Miao et al. (2012) identified boosting as significantly more accurate than bagging and Random Forest in the context of ecological zone classification. Some researchers have however discovered that the performance of boosting can be adversely affected by noise in the dataset. Bauer and Kohavi (1999), based on a study involving 14 datasets, concluded that although boosting in most cases outperformed bagging, it did not deal well with noise. Based on an extensive comparison containing 33 datasets, Dietterich (2000) concluded that in the case of noisy datasets, Random Forest outperformed both boosting and bagging. Ismail and Mutanga (2010) and Hamza and Larocque (2005) also identified Random Forests as the most accurate ensemble method when compared to both Boosted and Bagged Regression Trees for their operational (and thus noisy) applications. However, the effect of noise is dataset-dependent (Breiman 2001) and thus it remains impossible to predict which ensemble tree method will perform best for any given study. In our case, although we presume a considerable amount of noise in our data, the Boosted Regression Trees outperformed both the Bagged Regression Trees and the Random Forest.

Currently, classification and regression trees (CART) are the dominant techniques for MODIS (and Landsat TM) based classifications (Shao and Lunetta 2012). The standard MODIS global land cover classification is produced by a Boosted Regression Tree algorithm (Friedl et al. 2010). For our dataset, Boosted Regression Trees perform very well at the municipality level, but their accuracy remains below average at the pixel level. In general, the accuracy of all methods at the pixel level is rather low, with no method reaching a Nash-Sutcliffe index of 0.5. The RMSE for all methods exceeds 10 %, which is high considering

that the fractions within each pixel are assumed to sum to 100%. Comparable results are however obtained in other recent studies on sub-pixel land fraction estimation. Schwieder et al. (2014) obtained a minimum RMSE of 12% when using SVR and RF to estimate fractional shrub cover from simulated 30 m resolution EnMAP data. Shao and Lunetta (2012) obtained RMSE values at the pixel level between 24% en 50% for a MODIS-based sub-pixel land cover classification in Canada. One possible explanation for the average performance at the pixel level reported in this study is the large intra-class variability in phenology (and thus reflectance patterns). Although inductive methods may be better equipped to deal with this variability than endmember-based methods, it is likely to remain a major source of confusion between classes.

6.2 Training data and time constraints

Given that the cost of training data acquisition is often noted as a concern in remote sensing (Foody and Mathur 2006; Durbha, King, and Younan 2007; Chi, Feng, and Bruzzone 2008; Shao and Lunetta 2012), the ability to handle small data sets is an attractive feature for many applications, not in the least for land cover classification where reference data are often collected on the ground. A study conducted by Huang et al. (2002) suggests that the minimum number of samples for adequately training an algorithm may depend on the algorithm concerned, the number of input variables, the sampling strategy used to select the training samples and the size and heterogeneity of the study area. In our study, all algorithms were trained with the exact same training sets, thus excluding all influences except that of the algorithm itself.

For all methods the prediction accuracy increased with larger training sample sizes, which corresponds to the findings of other recent studies (Schwieder et al. 2014; Walton 2008; Shao and Lunetta 2012). Not all algorithms showed the same response to low numbers of observations. Boosted Regression Trees and Random Forests were capable of obtaining

fairly accurate predictions (municipality-level NS > 0.75) with as few as 10 training observations. The two support vector machine-based techniques, though often praised for their good performance with small training sets (Mountrakis, Im, and Ogole 2011; Chi, Feng, and Bruzzone 2008; Camps-Valls et al. 2004; Shao and Lunetta 2012; Song, Duan, and Jiang 2012) are in this study not competitive with BoRT and RF at the smallest training set sizes. In general, the accuracy gain between the larger training sample sizes is lower than that between the smaller ones.

The time needed for training may vary strongly with the processor type, thus our observations cannot be considered as absolute reference values about training time. Nevertheless, as all models were built and validated under equal conditions, our experimental set-up allows a valid relative comparison of the processing times required for model training.

6.3 Data quality

Although we tried to standardize all tests and neutralize random effects by working with ten replications for each model configuration, the effects of some sources of error could not be completely avoided. First, the quality of the input data is an important factor in any land cover classification. The disturbances resulting from cloud cover and atmospheric interference that affect the reflectance measured by the MODIS sensor are reduced by the constrained view angle maximum value (CV-MVC) temporal compositing step, as integrated into the MODIS preprocessing chain. In this study, an extra temporal smoothing step was included, to smoothen remaining peaks and dips in the temporal profile. We presume that the quality of the resulting NDVI time series is sufficiently high to spectrally distinguish the 16 land cover classes in our classification scheme based on their spectral profiles.

Another important consideration in land cover classification is the reliability of the reference data. The area-covering reference dataset used in this study was created through a spatial overlay of the level 2 CORINE land cover (CLC) 2006 dataset and an annually

updated regional dataset of agricultural parcels (EPR). The EPR is the most detailed and up-to-date agricultural land cover dataset available for Flanders. As the information about both parcel boundaries and crops cover are delivered by the parcel owners themselves, we have no reason to be particularly suspicious about its spatial and thematic accuracy. In the absence of definitive accuracy figures for CORINE Land Cover (CLC) 2006, the figures for CLC 2000 provide a basic indication. After its completion, the CLC 2000 database, the predecessor of CLC 2006, has been validated using LUCAS data (Kleeschulte and Büttner 2006). The European Land Use/Cover Area frame statistical Survey (LUCAS) is a project managed by Eurostat, the statistical office of the European Commission. A national validation for the Netherlands, a neighboring country of Flanders with comparably heterogeneous land cover patterns, revealed an overall accuracy of 95.8% for the level 2 CLC 2000 dataset (Hazeu, Dorland, and Schuiling 2008). We may thus presume that the level 2 accuracy of the CLC 2006 dataset will be situated somewhere within the range of 85-95%. As we use this data only to fill up the most stable land cover classes, we can safely assume a sufficient level of accuracy for its use as reference data in our modeling approach. The purpose of this study was to compare modeling approaches for use in an operational setting, which always come with some uncertainty about both the input and the reference data. This study thus provides an honest comparison of the selected approaches, with regard to their handling of these imperfect data sources.

6.4 Future research

The objective of this study was to provide some insight into the trade-offs that are encountered when deciding which method to use for a sub-pixel land cover classification in a spatially heterogeneous region. Optimizing a classification strategy however requires a broader perspective than simply selecting the most appropriate prediction model. Our lines for future research therefore consist of analyzing the effect of (i) input variable selection, (ii) the

classification scheme and (iii) imposing fractional abundance constraints on the performance of these machine learning methods. Fractional abundance constraints explicitly force an algorithm to deliver outputs that are nonnegative and sum to one. As that is exactly what area fractions constitute of – nonnegative values that sum to one – we expect the prediction accuracy to benefit from imposing the constraints.

7 Conclusion

The purpose of this paper was to evaluate the performance of six state-of-the-art machine learning methods for the specific task of sub-pixel land cover classification at the regional scale. The results confirm that classification accuracy alone does not suffice to allow an informed decision about the most appropriate method. Each method comes with its own set of meta-parameters which define the tediousness of the preprocessing phase a potential user has to perform. Moreover, the relative performance of the methods is largely influenced by the number of reference pixels available for training.

Our results indicate that Support Vector Machines and Boosted Regression Trees are able to generate the most accurate area fraction estimations at the municipality level, provided that unlimited preprocessing time and reference samples are available. In general, Support Vector Machines are less affected by (training) time constraints than Boosted Regression Trees. The impact of a training set size constraint on the other hand is more severe for Support Vector Machines than for Boosted Regression Trees. The overall rankings obtained from our performance scenarios lead us to conclude that, unless the number of training pixels is the major constraint, Support Vector Machines – both SVR and LS-SVM –should be favored for sub-pixel land cover classification in spatially heterogeneous regions.

Acknowledgements

This research was made possible through a PhD scholarship awarded by the Agency for Innovation by Science and Technology in Flanders (IWT) to Stien Heremans. The authors would also like to acknowledge the two anonymous reviewers for their remarks that helped us shape the final document.

References

Arora, M. K., and G. M. Foody. 1997. "Log-linear Modelling for the Evaluation of the Variables Affecting the Accuracy of Probabilistic, Fuzzy and Neural Network Classifications." *International Journal of Remote Sensing* 18 (4):785-798. doi: 10.1080/014311697218755.

Banfield, R. E., L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. 2007. "A Comparison of Decision Tree Ensemble Creation Techniques." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 29 (1):173-180.

Bauer, E., and R. Kohavi. 1999. "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants." *Machine Learning* 36 (1-2):105-139.

Beale, M., M. T. Hagan, and H. B. Demuth. 2012. "Neural Network Toolbox, User's Guide, MATLAB."

Benhadj, I., B. Duchemin, P. Maisongrande, V. Simonneaux, S. Khabba, and A.Chehbouni. 2012. "Automatic Unmixing of MODIS Multi-temporal Data for Inter-annual Monitoring of Land Use at a Regional Scale (Tensift, Morocco)." *International Journal of Remote Sensing* 33 (5):1325-1348. doi: 10.1080/01431161.2011.564220.

Berberoglu, S., O. Satir, and P. M. Atkinson. 2009. "Mapping Percentage Tree Cover from Envisat MERIS Data using Linear and Nonlinear Techniques." *International Journal of Remote Sensing* 30 (18):4747-4766. doi: 10.1080/01431160802660554.

Bocco, M., G. Ovando, S. Sayago, E. Willington, and S. Heredia. 2012. "Estimating Soybean Ground Cover from Satellite Images using Neural-networks Models." *International Journal of Remote Sensing* 33 (6):1717-1728. doi: 10.1080/01431161.2011.600347.

Bontemps, S., M. Herold, L. Kooistra, A. van Groenestijn, A. Hartley, O. Arino, I. Moreau, and P. Defourny. 2011. "Revisiting Land Cover Observations to Address the Needs of the Climate Modelling Community." *Biogeosciences Discussions* 8:7713-7740.

Breiman, L. 1996. "Bagging Predictors." *Machine Learning* 24 (2):123-140. doi: 10.1007/bf00058655.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1):5-32.

Brown, M., S. R. Gunn, and H. G. Lewis. 1999. "Support Vector Machines for Optimal Classification and Spectral Unmixing." *Ecological Modelling* 120 (2-3):167-179.

Büttner, G., G. Maucha, and B. Kosztra. 2011. "European Validation of Land Cover Changes in CLC2006 Project." In *31st EARSel Symposium 'Remote Sensing and Geoinformation not only for scientific cooperation'*, Czech Technical University, Prague, Czech Republic.

Camps-Valls, G., and L. Bruzzone. 2005. "Kernel-based Methods for Hyperspectral Image Classification." *Geoscience and Remote Sensing, IEEE Transactions on* 43 (6):1351-1362. doi: 10.1109/TGRS.2005.846154.

Camps-Valls, G., L. Gomez-Chova, J. Calpe-Maravilla, J. D. Martin-Guerrero, E. Soria-Olivas, L. Alonso-Chorda, and J. Moreno. 2004. "Robust Support Vector Method for Hyperspectral Data Classification and Knowledge Discovery." *Geoscience and Remote Sensing, IEEE Transactions on* 42 (7):1530-1542.

Carrão, H., A. Araújo, P. Gonçalves, and M. Caetano. 2010. "Multitemporal MERIS Images for Land-cover Mapping at a National Scale: a Case Study of Portugal." *International Journal of Remote Sensing* 31 (8):2063-2082. doi: 10.1080/01431160902942910.

Caruana, R., S. Lawrence, and L. Giles. 2001. "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping." *Advances in neural information processing systems*:402-408.

Caruana, R., and A. Niculescu-Mizil. 2006. "An Empirical Comparison of Supervised Learning Algorithms." *Proceedings of the 23rd international conference on Machine learning*, Pittsburgh, Pennsylvania.

Chang, C., and C. Lin. 2011. "LIBSVM: A Library for Support Vector Machines." *ACM Transactions on Intelligent Systems Technology* 2 (3):1-27. doi: 10.1145/1961189.1961199.

Chapman, D. S., A. Bonn, W. E. Kunin, and S. J. Cornell. 2010. "Random Forest Characterization of Upland Vegetation and Management Burning from Aerial Imagery." *Journal of Biogeography* 37 (1):37-46. doi: 10.1111/j.1365-2699.2009.02186.x.

Chen, J., P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh. 2004. "A Simple Method for Reconstructing a High-quality NDVI Time-series Data Set based on the Savitzky-Golay Filter." *Remote Sensing of Environment* 91 (3-4):332-344.

Cherkassky, V., and Y. Ma. 2004. "Practical Selection of SVM Parameters and Noise Estimation for SVM Regression." *Neural Networks* 17 (1):113-126.

Chi, M., R. Feng, and L. Bruzzone. 2008. "Classification of Hyperspectral Remote-sensing Data with Primal SVM for Small-sized Training Dataset Problem." *Advances in Space Research* 41 (11):1793-1799. doi: http://dx.doi.org/10.1016/j.asr.2008.02.012.

Clark, M. L., T. M. Aide, H. R. Grau, and G. Riner. 2010. "A Scalable Approach to Mapping Annual Land Cover at 250 m using MODIS Time Series Data: A case study in the Dry Chaco ecoregion of South America." *Remote Sensing of Environment* 114 (11):2816-2832. doi: http://dx.doi.org/10.1016/j.rse.2010.07.001.

Colditz, R. R., M. Schmidt, C. Conrad, M. C. Hansen, and S. Dech. 2011. "Land Cover Classification with Coarse Spatial Resolution Data to Derive Continuous and Discrete Maps for Complex Regions." *Remote Sensing of Environment* 115 (12):3264-3275. doi: 10.1016/j.rse.2011.07.010.

Cortés, G., M. Girotto, and S. A. Margulis. 2014. "Analysis of Sub-pixel Snow and Ice Extent over the Extratropical Andes using Spectral Unmixing of Historical Landsat Imagery." *Remote Sensing of Environment* 141:64-78. doi: http://dx.doi.org/10.1016/j.rse.2013.10.023.

De Brabanter, K., P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, and J. A. K. Suykens. 2010. "LS-SVMlab Toolbox User's Guide version 1.8." *International Report 10-146, ESAT-SISTA, KU Leuven (Leuven, Belgium)*.

DeFries, R. S., and J.C. Chan. 2000. "Multiple Criteria for Evaluating Machine Learning Algorithms for Land Cover Classification from Satellite Data." *Remote Sensing of Environment* 74 (3):503-515.

Dietterich, T. G. 2000. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision trees: Bagging, Boosting, and Randomization." *Machine Learning* 40 (2):139-157.

Dixon, B., and N. Candade. 2008. "Multispectral Landuse Classification using Neural Networks and Support Vector Machines: One or the Other, or Both?" *International Journal of Remote Sensing* 29 (4):1185-1206.
Dobreva, I. D., and A. G. Klein. 2011. "Fractional Snow Cover Mapping through Artificial Neural Network Analysis of MODIS Surface Reflectance." *Remote Sensing of Environment* 115 (12). doi: 10.1016/j.rse.2011.07.018.
Donlon, C., B. Berruti, A. Buongiorno, M. H. Ferreira, P. Féménias, J. Frerick, P. Goryl, U. Klein, H. Laur, C. Mavrocordatos, J. Nieke, H. Rebhan, B. Seitz, J. Stroede, and R. Sciarra. 2012. "The Global Monitoring for Environment and Security (GMES) Sentinel-3 mission." *Remote Sensing of Environment* 120:37-57.
Durbha, S. S., R. L. King, and N. H. Younan. 2007. "Support Vector Machines Regression for Retrieval of Leaf Area Index from Multiangle Imaging Spectroradiometer." *Remote Sensing of Environment* 107 (1-2):348-361.
Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*: Chapman & Hall.
Elith, J., J. R. Leathwick, and T. Hastie. 2008. "A Working Guide to Boosted Regression Trees." *Journal of Animal Ecology* 77 (4):802-813.
Faivre, R., and A. Fischer. 1997. "Predicting Crop Reflectances Using Satellite Data Observing Mixed Pixels." *Journal of Agricultural, Biological, and Environmental Statistics* 2 (1):87-107. doi: 10.2307/1400642.
Fan, F., and Y. Deng. 2014. "Enhancing Endmember Selection in Multiple Endmember Spectral Mixture Analysis (MESMA) for Urban Impervious Surface Area Mapping using Spectral Angle and Spectral Distance Parameters." *International Journal of Applied Earth Observation and Geoinformation* 33 (0):290-301. doi: http://dx.doi.org/10.1016/j.jag.2014.06.011.
Farook, A., M. B. Sivaraman, and S. Kesavaraj. 2013. "Sub-Pixel Classification of Multi-date Satellite Images for Accurate Change Detection in Pichavaram Mangroves, Tamilnadu, India." *Research & Reviews: Journal of Space Science & Technology* 2 (3).
Fisher, P. 1997. "The pixel: A Snare and a Delusion." *International Journal of Remote Sensing* 18 (3):679-685.
Foody, G. M. 1996. "Relating the Land-cover Composition of Mixed Pixels to Artificial Neural Network Classification Output." *Photogrammetric Engineering and Remote Sensing* 62 (5):491-499.
Foody, G. M., R. M. Lucas, P. J. Curran, and M. Honzak. 1997. "Non-linear Mixture Modelling without End-members using an Artificial Neural Network". *International Journal of Remote Sensing* 18 (4):937-953.
Foody, G. M., and A. Mathur. 2004. "Toward Intelligent Training of Supervised Image Classifications: Directing Training Data Acquisition for SVM Classification." *Remote Sensing of Environment* 93:107-117.
Foody, G. M., and A. Mathur. 2006. "The Use of Small Training Sets Containing Mixed Pixels for Accurate Hard Image Classification: Training on Mixed Spectral Responses for Classification by a SVM." *Remote Sensing of Environment* 103 (2):179-189.
Foody, G. M., M. B. McCulloch, and W. B. Yates. 1995. "Classification of Remotely Sensed Data by an Artificial Neural Network: Issues Related to Training Data Characteristics." *Photogrammetric Engineering & Remote Sensing* 61 (4):391-401.
Friedl, M. A., C. E. Brodley, and A. H. Strahler. 1999. "Maximizing Land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales." *Geoscience and Remote Sensing, IEEE Transactions on* 37 (2):969-977. doi: 10.1109/36.752215.

Friedl, M. A., D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang. 2010. "MODIS Collection 5 Global Land Cover: Algorithm Refinements and Characterization of New Datasets." *Remote Sensing of Environment* 114 (1):168-182. doi: http://dx.doi.org/10.1016/j.rse.2009.08.016.

Friedman, Jerome H. 2002. "Stochastic Gradient Boosting." *Computational Statistics & Data Analysis* 38 (4):367-378.

Gahegan, M. 2003. "Is Inductive Machine Learning just another Wild Goose (or Might it Lay the Golden Egg)?" *International Journal of Geographical Information Science* 17 (1):69-92. doi: 10.1080/13658810210157778.

Groisman, P. Y., and S. A. Bartalev. 2007. "Northern Eurasia Earth Science Partnership Initiative (NEESPI), Science Plan Overview." *Global and Planetary Change* 56:215-234.

Hamza, M., and D. Larocque. 2005. "An Empirical Comparison of Ensemble Methods based on Classification Trees." *Journal of Statistical Computation and Simulation* 75 (8):629-643. doi: 10.1080/00949650410001729472.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*: Springer.

Hazeu, G. W., G. J. van Dorland, and C. Schuiling. 2008. CLC2006 Land Cover Database of the Netherlands, Land Cover and Land Cover Changes 2000-2006. Copenhagen, Denmark: EEA.

Heremans, S., and J. Van Orshoven. 2011. "Effect of the Learning Algorithm on the Accuracy of Sub-pixel Land Use Classifications with Multilayer Perceptrons." *6th International Workshop on the Analysis of Multi-temporal Remote Sensing Images (Multi-Temp),* 2011, 12-14 July 2011.

Hird, Jennifer N., and Gregory J. McDermid. 2009. "Noise Reduction of NDVI Time Series: An Empirical Comparison of Selected Techniques." *Remote Sensing of Environment* 113 (1):248-258. doi: http://dx.doi.org/10.1016/j.rse.2008.09.003.

Hu, X., and Q. Weng. 2009. "Estimating Impervious Surfaces from Medium Spatial Resolution Imagery using the Self-organizing Map and Multi-layer Perceptron Neural Networks." *Remote Sensing of Environment* 113 (10):2089-2102. doi: DOI: 10.1016/j.rse.2009.05.014.

Huang, C., L. S. Davis, and J. R. G. Townshend. 2002. An Assessment of Support Vector Machines for Land Cover Classification. *International Journal of Remote Sensing* 23(4):725-749.

Huete, A., K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira. 2002. "Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Indices." *Remote Sensing of Environment* 83:195-213.

Ismail, R., and O. Mutanga. 2010. "A Comparison of Regression Tree Ensembles: Predicting Sirex Noctilio Induced Water Stress in Pinus Patula Forests of KwaZulu-Natal, South Africa." *International Journal of Applied Earth Observation and Geoinformation* 12, Supplement 1:S45-S51.

Jackson, Q., and D. A. Landgrebe. 2001. "An Adaptive Classifier Design for High-dimensional Data Analysis with a Limited Training Data Set." *Geoscience and Remote Sensing, IEEE Transactions on* 39 (12):2664-2679. doi: 10.1109/36.975001.

Jensen, J. 2005. *Introductory Digital Image Processing: A Remote Sensing Perspective*: Prentice Hall.

Jonsson, P., and L. Eklundh. 2004. "TIMESAT - a Program for Analyzing Time-series of Satellite Sensor Data." *Computers & Geosciences* 30 (8):833-845.

Kaptué T., A. T., J-L Roujean, and S. M. De Jong. 2011. "Comparison and Relative Quality Assessment of the GLC2000, GLOBCOVER, MODIS and ECOCLIMAP Land Cover

Sata Sets at the African Continental Scale." *International Journal of Applied Earth Observation and Geoinformation* 13 (2):207-219.
Kavzoglu, T., and I. Colkesen. 2009. "A Kernel Functions Analysis for Support Vector Machines for Land Cover Classification." *International Journal of Applied Earth Observation and Geoinformation* 11 (5):352-359.
Kavzoglu, T., and P. M. Mather. 2003. "The Use of Backpropagating Artificial Neural Networks in Land Cover Classification." *International Journal of Remote Sensing* 24 (23):4907-4938.
Kleeschulte, S., and G. Büttner. 2006. "European Land Cover Mapping - the CORINE Experience." *Proceedings of the North America Land Cover Summit* 2008:31-44.
Kohavi, R., D. Sommerfield, and J. Dougherty. 1997. "Data Mining using MLC++: a Machine Learning Library in C++." *International Journal on Artificial Intelligence Tools (Architectures, Languages, Algorithms)* 6 (4):537-66. doi: 10.1142/s021821309700027x.
Liaw, A. 2002. "Classification and Regression by randomForest." *R News* 2 (3):18.
Liu, W. G., and E. Y. Wu. 2005. "Comparison of Non-linear Mixture Models: Sub-pixel Classification." *Remote Sensing of Environment* 94 (2):145-154. doi: 10.1016/j.rse.2004.09.004.
Lu, D., and Q. Weng. 2007. "A Survey of Image Classification Methods and Techniques for Improving Classification Performance." *International Journal of Remote Sensing* 28 (5):823-870. doi: 10.1080/01431160600746456.
Lunetta, R. S., Y. Shao, J. Ediriwickrema, and J. G. Lyon. 2010. "Monitoring Agricultural Cropping Patterns across the Laurentian Great Lakes Basin using MODIS-NDVI Data." *International Journal of Applied Earth Observation and Geoinformation* 12 (2):81-88. doi: http://dx.doi.org/10.1016/j.jag.2009.11.005.
Matsuoka, M., T. Hayasaka, Y. Fukushima, and Y. Honda. 2007. "Land Cover in East Asia Classified using Terra MODIS and DMSP OLS Products." *International Journal of Remote Sensing* 28 (2):221-248. doi: 10.1080/01431160600675911.
Miao, X., J. S. Heaton, S. Zheng, D. A. Charlet, and H. Liu. 2012. "Applying Tree-based Ensemble Algorithms to the Classification of Ecological Zones using Multi-temporal Multi-source Remote-sensing Data." *International Journal of Remote Sensing* 33 (6):1823-1849. doi: 10.1080/01431161.2011.602651.
Mountrakis, G., J. Im, and C. Ogole. 2011. "Support Vector Machines in Remote Sensing: A review." *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (3):247-259.
Nash, J. E., and J. V. Sutcliffe. 1970. "River Flow Forecasting through Conceptual Models Part I — A Discussion of Principles." *Journal of Hydrology* 10 (3):282-290. doi: http://dx.doi.org/10.1016/0022-1694(70)90255-6.
Pal, M., and P. M. Mather. 2003. "An Assessment of the Effectiveness of Decision Tree Methods for Land Cover Classification." *Remote Sensing of Environment* 86 (4):554-565.
Peters, A., T.n Hothorn, and M. T. Hothorn. 2012. Package 'ipred'. edited by The R Foundation for Statistical Computing.
Peters, J., B. De Baets, N. E. C. Verhoest, R. Samson, S. Degroeve, P. De Becker, and W. Huybrechts. 2007. "Random Forests as a Tool for Ecohydrological Distribution modelling." *Ecological Modelling* 207 (2-4):304-318.
Poelmans, Lien, and Anton Van Rompaey. 2009. "Detecting and Modelling Spatial Patterns of Urban Sprawl in Highly Fragmented Areas: A Case Study in the Flanders–Brussels Region." *Landscape and Urban Planning* 93 (1):10-19. doi: http://dx.doi.org/10.1016/j.landurbplan.2009.05.018.

Prasad, A., L. Iverson, and Andy Liaw. 2006. "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction." *Ecosystems* 9 (2):181-199. doi: 10.1007/s10021-005-0054-1.

Reschke, J., and C. Hüttich. 2014. "Continuous Field Mapping of Mediterranean Wetlands using Sub-pixel Spectral Signatures and Multi-Temporal Landsat Data." *International Journal of Applied Earth Observation and Geoinformation* 28 (0):220-229. doi: http://dx.doi.org/10.1016/j.jag.2013.12.014.

Ridgeway, G. 2007. Generalized Boosted Models: a Guide to the gbm Package.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez. 2012. "An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 67:93-104.

Rogan, J., J. Franklin, D. Stow, J. Miller, C. Woodcock, and D. Roberts. 2008. "Mapping Land-cover Modifications over Large Areas: A Comparison of Machine Learning Algorithms." *Remote Sensing of Environment* 112 (5):2272-2283.

Roosta, H., and M. R. Saradjian. 2007. "Sub-pixel Classification of MODIS Images." *Proceedings of the 6th WSEAS International Conference on Non-linear analysis, Non-linear Systems and Chaos*, Arcachon, France.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. "Learning Internal Representations by Error Propagation." In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1*, 318-362. MIT Press.

Schwieder, M., P. J. Leitão, S. Suess, C. Senf, and P. Hostert. 2014. "Estimating Fractional Shrub Cover Using Simulated EnMAP Data: A Comparison of Three Machine Learning Regression Techniques." *Remote Sensing* 6 (4):3427-3445.

Shao, Y., and R. S. Lunetta. 2011. "Sub-Pixel Mapping of Tree Canopy, Impervious Surfaces, and Cropland in the Laurentian Great Lakes Basin Using MODIS Time-Series Data." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4 (2):336-347. doi: 10.1109/jstars.2010.2062173.

Shao, Y., and R. S. Lunetta. 2012. "Comparison of Support Vector Machine, Neural Network, and CART Algorithms for the Land-cover Classification using Limited Training Data Points." *ISPRS Journal of Photogrammetry and Remote Sensing* 70 (0):78-87. doi: http://dx.doi.org/10.1016/j.isprsjprs.2012.04.001.

Shao, Y., R. S. Lunetta, J. Ediriwickrema, and J. Iiames. 2009. "Mapping Cropland and Major Crop Types across the Great Lakes Basin using MODIS-NDVI Data." *Photogrammetric Engineering & Remote Sensing* 75(1):73-84.

Shataee, S., S. Kalbi, A. Fallah, and D. Pelz. 2012. "Forest Attribute Imputation using Machine-learning Methods and ASTER Data: Comparison of k-NN, SVR and Random Forest Regression Algorithms." *International Journal of Remote Sensing* 33 (19):6254-6280. doi: 10.1080/01431161.2012.682661.

Smola, A. J., and B. Scholkopf. 2004. "A Tutorial on Support Vector Regression." *Statistics and Computing* 14 (3):199-222.

Song, X., Z. Duan, and X. Jiang. 2011. "Comparison of Artificial Neural Networks and Support Vector Machine Classifiers for Land Cover Classification in Northern China using a SPOT-5 HRG Image." *International Journal of Remote Sensing* 33 (10):3301-3320. doi: 10.1080/01431161.2011.568531.

Strobl, C., J. Malley, and G. Tutz. 2009. "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests." *Psychological Methods* 14 (4):323-348. doi: 10.1037/a0016973.

Sutton, Clifton D. 2005. "Classification and Regression Trees, Bagging, and Boosting." In *Handbook of Statistics*, 303-329. Elsevier.

Suykens, J A K, T Van Gestel, J De Brabanter, B De Moor, and J. Vandewalle. 2002. *Least Squares Support Vector Machines*: World Scientific Pub.Co., Singapore.

Szuster, Brian W., Qi Chen, and Michael Borger. 2011. "A Comparison of Classification Techniques to Support Land Cover and Land Use Analysis in Tropical Coastal Zones." *Applied Geography* 31 (2):525-532.

Van Daele, T. W. Van Reeth, M. Dumortier, and J. Peymen. 2010. "Biodiversity indicators 2010. The State of Nature in Flanders (Belgium)." *Research Institute for Nature and Forest, Brussels. INBO*.

Vapnik, V. N. 2005. *The Nature of Statistical Learning Theory*.

Verbeiren, S., H. Eerens, I. Piccard, I. Bauwens, and J. Van Orshoven. 2008. "Sub-pixel Classification of SPOT-VEGETATION time series for the assessment of regional crop areas in Belgium." *International Journal of Applied Earth Observation and Geoinformation* 10 (4):486-497.

Verburg, P. H., K. Neumann, and L. Nol. 2011. "Challenges in using Land Use and Land Cover Data for Global Change Studies." *Global Change Biology* 17 (2):974-989. doi: 10.1111/j.1365-2486.2010.02307.x.

Verrelst, J., J. Munoz, L. Alonso, J. Delegido, J. P. Rivera, G. Camps-Valls, and J. Moreno. 2012. "Machine Learning Regression Algorithms for Biophysical Parameter Retrieval: Opportunities for Sentinel-2 and -3." *Remote Sensing of Environment* 118:127-139.

Walton, J. T. 2008. "Subpixel Urban Land Cover Estimation: Comparing Cubist, Random Forests, and Support Vector Regression." *Photogrammetric Engineering and Remote Sensing* 74 (10):1213-1222.

Wang, H., Y. Shao, and L. M. Kennedy. 2014. "Temporal Generalization of Sub-pixel Vegetation Mapping with Multiple Machine Learning and Atmospheric Correction Algorithms." *International Journal of Remote Sensing* 35 (20):7118-7135. doi: 10.1080/01431161.2014.965288.

Wardlow, B. D., S. L. Egbert, and J. H. Kastens. 2007. "Analysis of Time-series MODIS 250 m Vegetation Index Data for Crop Classification in the U.S. Central Great Plains." *Remote Sensing of Environment* 108 (3):290-310. doi: DOI: 10.1016/j.rse.2006.11.021.

Weng, Q. 2012. "Remote Sensing of Impervious Surfaces in the Urban Areas: Requirements, Methods, and Trends." *Remote Sensing of Environment* 117:34-49. doi: http://dx.doi.org/10.1016/j.rse.2011.02.030.

Wilamowski, B. M., S. Iplikci, O. Kaynak, and M. O. Efe. 2001. "An Algorithm for Fast Convergence in Training Neural Networks." *Proceedings of the International Joint Conference on Neural Networks, 2001*.

Zhang, Y., H. Zhang, and H. Lin. 2014. "Improving the Impervious Surface Estimation with Combined Use of Optical and SAR Remote Sensing Images." *Remote Sensing of Environment* 141:155-167. doi: http://dx.doi.org/10.1016/j.rse.2013.10.028.

Zhuang, X., B. A. Engel, D. F. Lozano-Garcia, R. N. Fernandez, and C. J. Johannsen. 1994. "Optimization of Training Data Required for Neuro-classification." *International Journal of Remote Sensing* 15 (16):3271-3277. doi: 10.1080/01431169408954326.